Sparse Multiple Kernel Learning with Geometric Convergence Rate

Rong Jin RONGJIN@CSE.MSU.EDU
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI, 48824, USA
Tianbao Yang TYANG@GE.COM
Machine Learning Lab
GE Global Research
San Ramon, CA 94583,, USA
Mehrdad Mahdavi MAHDAVIM@CSE.MSU.EDU
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI, 48824, USA

Abstract

In this paper, we study the problem of sparse multiple kernel learning (MKL), where the goal is to efficiently learn a combination of a fixed small number of kernels from a large pool that could lead to a kernel classifier with a small prediction error. We develop an efficient algorithm based on the greedy coordinate descent algorithm, that is able to achieve a geometric convergence rate under appropriate conditions. The convergence rate is achieved by measuring the size of functional gradients by an empirical ℓ_2 norm that depends on the empirical data distribution. This is in contrast to previous algorithms that use a functional norm to measure the size of gradients, which is independent from the data samples. We also establish a generalization error bound of the learned sparse kernel classifier using the technique of local Rademacher complexity.

Keywords: kernel methods, multiple kernel learning, greedy coordinate descent, generalization bound

1. Introduction

Kernel methods have been studied extensively, thanks to their empirical success in a variety of applications. Examples of kernel methods include support vector machines (SVMs), kernel ridge regression, kernel clustering, kernel PCA, and many others. It is well known that the choice of kernel function can be crucial to the success of kernel methods. Although, in principle kernel can be chosen by standard model selection methods such as cross validation, the high computational cost makes it unattractive. Over the past decade, significant progress has been made to efficiently learn an appropriate kernel for a given task.

Among the many approaches developed for kernel learning, recent studies have been focused predominately on multiple kernel learning (MKL) algorithms. Given a collection of kernels, the objective of MKL is to learn a combination of multiple kernel clas-

sifiers, one for each kernel function, from the training examples that results in small prediction error. Many computational algorithms have been developed for multiple kernel learning (Lanckriet et al., 2004; Argyriou et al., 2005; Bach, 2008; Argyriou et al., 2006; Lewis et al., 2006; Micchelli and Pontil, 2005; Ong et al., 2005; Bach et al., 2004; Rakotomamonjy et al., 2008; Sonnenburg et al., 2006; Xu et al., 2008; Suzuki and Tomioka, 2011). The analysis of generalization error bound for MKL has been developed in several studies (Hussain and Shawe-Taylor, 2011; Ying and Zhou, 2007; Cortes et al., 2009, 2010; Bousquet and Herrmann, 2003; Srebro and Ben-david, 2006; Ying and Campbell, 2009), aiming to bound the additional error arising from optimizing the combination of multiple kernels. These studies have shown that MKL can be effective even when the number of kernels to be combined is very large. For instance, the generalization error bound from learning a combination of m different kernels, will only deteriorate by a factor of $\log m$ when the sum of kernel combination weights is bounded.

Despite the encouraging results, one problem with MKL is that the resulting classifier can be a combination of many kernel classifiers, leading to a high computational cost in testing. We address this challenge by developing efficient algorithms and theories for sparse multiple kernel learning. The objective of sparse MKL is to learn a sparse combination of multiple kernel classifiers involving no more than d kernels, where $d \ll m$ is a predefined constant.

We develop a simple algorithm for learning such a sparse combination of multiple kernel classifiers, and present the analysis bounding the generalization performance of the learned kernel classifier. Our algorithm is an iterative algorithm based on the greedy coordinate descent algorithm (Shalev-Shwartz et al., 2010; Nesterov, 2010; Yun et al., 2011). To generate a sparse MKL solution involving no more than d kernels, at each iteration, our algorithm adds to the existing pool the kernel with the largest gradient. The size of gradients is measured by an empirical ℓ_2 norm that depends on the training examples. Under appropriate condition, the proposed approach is able to achieve a geometric convergence rate. To the best of our knowledge, this is the first algorithm for sparse MKL that achieves a geometric convergence rate.

Although several algorithms have been developed for sparse MKL by exploring different forms of regularization (Vishwanathan et al., 2010; Kloft et al., 2009; Orabona and Jie, 2011), none of them are able to establish the generalization error bound for a MKL solution involved a fixed number (i.e., d) of kernels. We also note that our work differs from the studies on the sparsity of MKL (Koltchinskii and Yuan, 2008, 2010) which focus on bounding the sparsity of combination weights for kernels and do not address our problem directly.

The most related work to this study is (Sindhwani and Lozano, 2011), where a group orthogonal matching pursuit (GOMP) algorithm is applied to learn a sparse combination of kernel classifiers with exactly d kernels. Unlike previous formulations for sparse MKL that use ℓ_1 regularization (i.e. $\sum_j \|f_j\|_{\mathcal{H}_{\kappa_j}}$), the authors propose to use ℓ_2^2 regularization (i.e. $\sum_j \|f_j\|^2_{\mathcal{H}_{\kappa_j}}$) together with a sparsity constraint (i.e. ℓ_0 constraint) for sparse MKL. Although they did not present a convergence analysis for the proposed algorithm except for a sparse recovery analysis, we can apply the analysis in (Shalev-Shwartz et al., 2010) for smooth functions to their algorithm to obtain a $O(1/d)$ convergence rate. The group orthogonal matching pursuit algorithm is similar to the greedy coordinate descent algorithm used in this study except that we measure the size of gradients by an empirical ℓ_2 norm

while it is measured by a functional norm in (Sindhwani and Lozano, 2011). It is this difference that leads to a geometric convergence rate for the proposed algorithm which is a significant improvement over the rate of $O(1/d)$.

Outline of contributions. The following contributions are made in this paper:

- We present a baseline algorithm, based on the greedy coordinate descent method, that achieves $O(1/d)$ convergence rate when using ℓ_1 norm functional regularizer.
- We introduce an empirical ℓ_2 norm to measure the size of functional gradients in the application of greedy coordinate descent algorithm to sparse MKL, and achieve a geometric convergence rate under appropriate conditions.
- We study the generalization performance of the proposed algorithm. Specifically, we derive an upper bound on the generalization performance of learned classifier using local Rademacher technique that has a additive term of $O\left(d\sqrt{\ln m/N}\right)$, which matches the existing bounds in their dependence on m (i.e., the number of kernel functions) and N (i.e., the number of training samples).

Our paper is organized as follows. In the next section we formally introduce the problem of sparse MKL. In section 3 we present our baseline algorithm with its convergence analysis. Section 4 introduces the main algorithm proposed in this paper with analysis of its convergence rate and generalization bound. We wrap up in Section 5 with a discussion of possible directions for the future work.

2. Problem Setting: Sparse Multiple Kernel Learning (MKL)

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, N\}$ be a collection of training examples, where $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \{-1, +1\}$, and let $\{\kappa_j(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}, j \in [m]\}$ be a collection of reproducing kernels to be combined, where $[m]$ denotes the set $\{1, \cdots, m\}$. Let $\{\mathcal{H}_j, j \in [m]\}$ be the associated Reproducing Kernel Hilbert Spaces (RKHS). We denote by $\mathbf{y} = (y_1, \ldots, y_N)^\top$ the outputs for all the instances in $\mathcal{D}$. For the convenience of analysis, we assume $\kappa_j(\mathbf{x}, \mathbf{x}) \leq 1$ for any $\mathbf{x} \in \mathcal{X}$ and any $j \in [m]$. The goal of MKL is to learn a function $f = \sum_{j=1}^m f_j$, where $f_j \in \mathcal{H}_j, j \in [m]$, that has a small generalization error. A common approach for MKL is to learn the combination of kernel classifiers by solving the following optimization problem (Micchelli and Pontil, 2005)

$$\min_{f \in \mathcal{H}} \quad \mathcal{L}(f) = \frac{1}{N} \sum_{i=1}^N \ell(f(\mathbf{x}_i), y_i) + \lambda \sum_{j=1}^m \|f_j\|_{\mathcal{H}_j}, \tag{1}$$

where $\mathcal{H} = \{f = \sum_{j=1}^m f_j : f_j \in \mathcal{H}_j\}$, and $\ell(z, y) = (z - y)^2/2$ is a square loss [1]. In this study, we assume that the number of kernels m is very large (could be larger than the number of training examples N), and our objective is to learn a combination of kernel classifiers involving no more than d kernels, where $d \ll m$ is a predefined constant. For the convenience of discussion, we define by $\mathcal{E}_N(f) = \frac{1}{N} \sum_{i=1}^N \ell(f(\mathbf{x}_i), y_i)$ the empirical loss for

1. Although we restrict our discussion to square loss, it is straightforward to extend our result to the quadratic-type loss function defined in (Koltchinskii and Yuan, 2010)

kernel classifier f, by $\|f\| = \sum_{j=1}^{m} \|f_j\|_{\mathcal{H}_j}$ the norm of a combined kernel classifier f, and by $J(f) = \{j \in [m] : f_j \neq 0\}$ the subset of non-zero kernel classifiers used to construct f. Finally, we define f^* the optimal solution to (1), i.e.,

$$f^* = \arg\min_{f \in \mathcal{H}} \mathcal{L}(f). \tag{2}$$

Note that according to (Micchelli and Pontil, 2005), the problem in (1) is equivalent to the following optimization problem

$$\min_{\gamma \in \mathbb{R}_+^m, \gamma^\top \mathbf{1} \leq 1} \min_{f \in \mathcal{H}_\gamma} \frac{\lambda'}{2} \|f\|_{\mathcal{H}_\gamma}^2 + \frac{1}{N} \sum_{i=1}^{N} (f(\mathbf{x}_i) - y_i)^2, \tag{3}$$

where $\lambda' > 0$ is an appropriately chosen parameter depending on λ in (1), and $\mathcal{H}_\gamma$ is a RKHS endowed with a combined kernel function $\kappa(\cdot, \cdot; \gamma) = \sum_{i=1}^{m} \gamma_i \kappa_j(\cdot, \cdot)$. It is not difficult to show that γ_j computed in (3) is proportional to $\|f_j\|_{\mathcal{H}_j}$ computed from (1). As a result, choosing the kernel classifiers with the largest functional norm in (1) is equivalent to choosing the kernels with the largest weights γ_j in (3).

3. Warmup: A Greedy Coordinate Descent Algorithm for Sparse MKL

A straightforward approach for sparse MKL is a two-stage scheme: it first learns a combination of all m kernels by solving the problem in (1) and then only keeps the d most "important" kernel classifiers f_j in the combination. To select the most important kernel classifiers, a simple approach is to choose the kernel classifiers with the largest functional norm $\|f_i\|_{\mathcal{H}_{\kappa_i}}$, because $\|f_i\|_{\mathcal{H}_{\kappa_i}}$ is proportional to the combination weight γ_i in (3). It is however easy to construct a counter example to show that the two-stage scheme fails to find the best kernel. In particular, we will show that for two cases that have the same sets of unique kernels, the two-stage scheme chooses different kernels. In the first case, we have two kernel functions $\kappa_1(\cdot, \cdot)$ and $\kappa_2(\cdot, \cdot)$. Using multiple kernel learning, we can learn the weights for both kernels. Let the learned weights be 0.8 for $\kappa_1(\cdot, \cdot)$ and 0.2 for $\kappa_2(\cdot, \cdot)$. According to the two-stage approach, we will select kernel $\kappa_1(\cdot, \cdot)$. In the second case, we have 10 identical copies of $\kappa_1(\cdot, \cdot)$ and one copy of κ_2. Since both cases share the same set of unique kernels, we expect the same kernel to be selected by the two-stage approach. However, based on the symmetric argument, it is straightforward to show that the weight for $\kappa_2(\cdot, \cdot)$ remains unchanged while the weights for the copies of $\kappa_1(\cdot, \cdot)$ are reduced to 0.08. As a result, the two-stage approach selects kernel $\kappa_2(\cdot, \cdot)$ for the second case, a different kernel from the first case. Another problem with this two-stage approach is its high computational complexity since it requires solving an optimization problem involved all kernel functions, even including the ones that are totally irrelevant to the target prediction task.

As the first step, we present a baseline algorithm that extends the greedy coordinate descent algorithm (Shalev-Shwartz et al., 2010) to solve the ℓ_1 regularized MKL in (1) and achieves a $O(1/d)$ convergence rate. The basic steps are shown in Algorithm 1. At each iteration k, Algorithm 1 selects the kernel with the largest gradient measured by its functional norm, denoted by j_k, and expands the set of selected kernels $\mathcal{S}_k$ to $\mathcal{S}_{k+1}$ by including j_k. It then searches for the optimal combination of kernels in the set $\mathcal{S}_{k+1}$ that

Algorithm 1 A Greedy Coordinate Descent Approach for Sparse MKL with ℓ_1 Regularization

1: **Input**: $\lambda > 0$: regularization parameter, d: the number of selected kernels
2: **Initialization**: $f_j^0 = 0, j \in [m]$ and $\mathcal{S}_0 = \emptyset$.
3: **for** $k = 1, \ldots, d$ **do**
4: $\quad j_k = \arg\max_{j \in [m]} \left\| \nabla_j \mathcal{E}_N(f^{k-1}) \right\|_{\mathcal{H}_j}$
5: $\quad$ Exist the loop if $\left\| \nabla_{j_k} \mathcal{E}_N(f^{k-1}) \right\|_{\mathcal{H}_{j_k}} \leq \lambda$
6: $\quad \mathcal{S}_k = \mathcal{S}_{k-1} \cup \{j_k\}$
7: $\quad$ Update the kernel classifier by solving the following optimization problem

$$f^k = \arg\min_{J(f) = \mathcal{S}_k} \mathcal{L}(\mathbf{w}) = \lambda \|f\| + \mathcal{E}_N(f) \tag{4}$$

8: **end for**
9: **Output** $f = f^{k-1}$

minimizes the objective function $\mathcal{L}(f)$. Note that although the objective in (1) is non-smooth due to the non-smooth regularization term $\sum_{j=1}^m \|f_j\|_{\mathcal{H}_j}$, we are still able to obtain a $O(1/d)$ convergence rate as shown in Theorem 1. The magic lies in step 4, where instead of choosing the coordinate with the largest gradient with respect to the objective function $\mathcal{L}(f)$, we choose the coordinate with the largest gradient with respect to $\mathcal{E}_N(f)$, the smooth part in the objective function, i.e.

$$\|\nabla_j \mathcal{E}_N(f)\|_{\mathcal{H}_{\kappa_j}} = \left\| \frac{1}{N} \sum_{i=1}^N \ell'(f(\mathbf{x}_i), y_i) \kappa_j(\mathbf{x}_i, \cdot) \right\|_{\mathcal{H}_{\kappa_j}}.$$

On the other hand, in step 7, we update the multiple kernel classifier by solving the ℓ_1 regularized MKL. It is this special design that makes it possible to achieve $O(1/d)$ convergence rate even for the non-smooth objective function in (1). We finally note that Algorithm 1 is similar in spirit to the GOMP based approach (Sindhwani and Lozano, 2011) and share the same convergence rate. The main difference is that we directly solve the ℓ_1 regularized MKL in (1) while in (Sindhwani and Lozano, 2011), a ℓ_2^2 regularizer is used and the sparsity is enforced through a constraint based on the ℓ_0 norm.

The following theorem shows the performance guarantee of the solution obtained by Algorithm 1 where its proof is given in Appendix A.

Theorem 1 *Let f be the solution output from Algorithm 1. If f is obtained by exiting from the middle of the loop, we have $\mathcal{L}(f) = \mathcal{L}(f^*)$. Otherwise, we have*

$$\mathcal{E}_N(f) + \lambda \|f\| \leq \mathcal{E}_N(f^*) + \lambda \|f^*\| + \frac{2}{d-1} \|f^*\|^2.$$

It should be emphasized that although the analysis in (Shalev-Shwartz et al., 2010) shows that the greedy coordinate descent approaches enjoy a geometric convergence rate when the objective function is both strongly convex and smooth in its variables, it can not be applied to our problem directly. This is because although the loss function $\ell(z, y)$

used in the regression is both strongly convex and smooth in the argument y, it is not strongly convex in $\{f_j\}_{j=1}^m$ because the prediction is given by $\sum_{j=1}^m f_j(\mathbf{x})$. In next section, we present another approach for sparse MKL, based on greedy coordinate descent, that is able to achieve a geometric convergence rate under appropriate conditions.

4. A Geometrically Convergent Algorithm for Sparse MKL

In this section, we present an algorithm for sparse MKL that can achieve a geometric convergence rate under appropriate conditions.

We first argue that selecting kernel classifiers based on their functional norm may not necessarily be the best idea. This is because in order to ensure a removed kernel classifier f_j to have a small impact on the overall regression error, we should be mostly concerned with $\mathrm{E}[|f_j(\mathbf{x})|^2]$, instead of $\|f_j\|_{\mathcal{H}_j}$. To see this, we bound $\mathrm{E}[|f(\mathbf{x})-y|^2]-\mathrm{E}[|f(\mathbf{x})-f_j(\mathbf{x})-y|^2]$, which measures the impact of removing f_j from f

$$\begin{aligned}\mathrm{E}[|f(\mathbf{x})-f_j(\mathbf{x})-y|^2]-\mathrm{E}[|f(\mathbf{x})-y|^2] &= \mathrm{E}[|f_j(\mathbf{x})|^2]-2\mathrm{E}[f_j(\mathbf{x})(f(\mathbf{x})-y)]\\ &\le \mathrm{E}[|f_j(\mathbf{x})|^2]+2\sqrt{\mathrm{E}[|f_j(\mathbf{x})|^2]}\sqrt{\mathrm{E}[|f(\mathbf{x})-y|^2]}.\end{aligned}$$

Although $\|f_j\|_{\mathcal{H}_j}\ge\|f_j\|_\infty\ge\sqrt{\mathrm{E}[|f_j(\mathbf{x})|^2]}$, there could be a significant gap between $\|f_j\|_{\mathcal{H}_j}$ and $\sqrt{\mathrm{E}[|f_j(\mathbf{x})|^2]}$ (Smale and Zhou, 2007), making it possible for the functional norm based criterion to *remove* the kernels that are *important* in the final prediction.

Based on the above discussion, we propose to measure the size of kernel classifiers f_j by its ℓ_2 norm, i.e., $\sqrt{\mathrm{E}|f_j(\mathbf{x})|^2}$. Since the distribution of $\mathbf{x}$ is unavailable, we introduce the empirical counterpart of $\sqrt{\mathrm{E}|f_j(\mathbf{x})|^2}$, called empirical ℓ_2 norm and denoted by $\|f_j\|_{\ell_2(\mathcal{D})}$. Given $f_j=\sum_{i=1}^N\alpha_{ji}\kappa_j(\mathbf{x}_i,\cdot)$, its $\ell_2(\mathcal{D})$ norm is computed as

$$\|f_j\|_{\ell_2(\mathcal{D})}=\sqrt{\frac{1}{N}\sum_{a=1}^N f_j^2(\mathbf{x}_a)}=\sqrt{\frac{1}{N}\sum_{a=1}^N\left(\sum_{b=1}^N\alpha_{jb}\kappa_j(\mathbf{x}_b,\mathbf{x}_a)\right)^2}=\frac{1}{\sqrt{N}}\|K_j\alpha_j\|_2, \tag{5}$$

where $K_j=[\kappa_j(\mathbf{x}_a,\mathbf{x}_b)]_{N\times N}$ is the kernel matrix for $\kappa_j(\cdot,\cdot)$, and $\alpha_j=(\alpha_{j1},\cdots,\alpha_{jN})^\top$. For the purpose of our analysis, we also define an empirical ℓ_2 norm for the combined classifier $f=\sum_{j=1}^m f_j=\sum_{j=1}^m\sum_{i=1}^N\alpha_{ji}\kappa_j(\mathbf{x}_i,\cdot)$ as

$$\|f\|_{\ell_2(\mathcal{D})}=\sqrt{\frac{1}{N}\sum_{i=1}^N f^2(\mathbf{x}_i)}=\frac{1}{\sqrt{N}}\left\|\sum_{j\in[m]}K_j\alpha_j\right\|_2. \tag{6}$$

One way to exploit the empirical ℓ_2 norm for sparse MKL is to incorporate it into (1) as part of the regularization, leading to a mixture regularizer that is consisted of both $\|f_j\|_{\mathcal{H}_j}$ and $\|f_j\|_{\ell_2(\mathcal{D})}$. A similar formulation is suggested in (Koltchinskii and Yuan, 2010). It is however unclear as how to efficiently solve the related optimization problem to achieve a convergence rate better than $O(1/d)$. Instead, we will use the empirical $\ell_2(\mathcal{D})$ norm to measure the size of gradients when performing greedy coordinate descent optimization. Our

Algorithm 2 A $\ell_2(\mathcal{D})$ Norm based Greedy Coordinate Descent Approach for Sparse MKL

Input: $\lambda > 0$: regularization parameter, d: the number of selected kernels
Initialization: $f_j^0 = 0, j \in [m]$ and $\mathcal{S}_0 = \emptyset$.
for $k = 1, \ldots, d$ **do**
  $j_k = \arg\max_{j\in[m]} \left\|\nabla_j \mathcal{E}_N(f^{k-1})\right\|_{\ell_2(\mathcal{D})}$
  Update the kernel classifier as

$$f^k = f^{k-1} - f_{j_k}, \text{ where } f_{j_k} = \frac{1}{N}\sum_{i=1}^{N} a_i^k \kappa_{j_k}(\mathbf{x}_i, \cdot) \tag{7}$$

  where $\mathbf{a}^k = \begin{pmatrix} a_1^k \\ \ldots \\ a_N^k \end{pmatrix}$ is the projection of $\boldsymbol{\ell}'(f^{k-1}) = \begin{pmatrix} \ell'(f^{k-1}(\mathbf{x}_1), y_1) \\ \ldots \\ \ell'\left(f^{k-1}(\mathbf{x}_N), y_N\right) \end{pmatrix}$ into the space spanned by the column vectors of the kernel matrix $K_{j_k} = [\kappa_{j_k}(\mathbf{x}_a, \mathbf{x}_b)]_{N\times N}$.
end for
Output $f = f^d$

analysis in subsection 4.1 shows that this modification to Algorithm 1, together with other changes, will result in a geometric convergence rate under appropriate conditions, i.e.

$$\mathcal{E}_N(f) - \min_{f\in\mathcal{H}} \mathcal{E}_N(f) \leq O(\max(0, (1-\tau)^d)),$$

where the value of τ will be determined by analysis.

Algorithm 2 gives the basic steps of the new approach for sparse MKL. Similar to Algorithm 1, at each iteration, Algorithm 2 chooses the kernel with the largest gradient and updates the kernel classifier based on the gradient with respect to the selected kernel. The key difference between these two algorithms is how to measure the size of the gradients. In Algorithm 1, the size of gradient $\nabla_j \mathcal{E}_N(f^k)$ is measured by its functional norm, while Algorithm 2 measures the size of gradient by $\ell_2(\mathcal{D})$ norm of $\nabla_j \mathcal{E}_N(f^k)$. In addition, Algorithm 2 follows the idea of gradient descent for updating the kernel classifier f^k and does not require solving any optimization problem. However, unlike the standard gradient descent algorithm that updates the classifier directly using the gradient, Algorithm 2 projects the coefficients of $\nabla_{j_k} \mathcal{E}_N(f^{k-1})$ into the subspace spanned by the column vectors in K_j before using it for updating. This step is critical for the correctness of the algorithm.

4.1. Convergence Analysis

To analyze the performance of Algorithm 2, we assume there exists a sparse MKL solution that achieves a small regression error. More specifically, we slightly abuse our notation by redefining f^* as the optimal kernel classifier that minimizes the empirical loss $\mathcal{E}_N(f)$, $\hat{f}$ as the optimal kernel classifier that minimizes the empirical loss using no more than d kernels, and ε^* be the difference in the empirical loss between $\hat{f}$ and f^*, i.e.,

$$f^* = \arg\min_{f\in\mathcal{H}} \mathcal{E}_N(f), \quad \hat{f} = \arg\min_{f\in\mathcal{H},\, |J(f)|\leq d} \mathcal{E}_N(f), \quad \varepsilon^* = \mathcal{E}_N(\hat{f}) - \mathcal{E}_N(f^*). \tag{8}$$

We assume ε^* is small, implying that the optimal solution f^* can be well approximated by a function involved no more than d kernels.

In order to state our result, we need to characterize the relationship among different kernel matrices. In (Koltchinskii, 2011), the author defines quantity $\beta(b, J, H)$ to capture the geometric relationship for a set of vectors $H = (\mathbf{h}_1, \ldots, \mathbf{h}_m) \in \mathbb{R}^{N\times m}$, i.e.,

$$\beta(b, J, H) = \inf \left\{ \beta > 0 : \sum_{j\in J} \lambda_j^2 \leq \beta^2 \left\| \sum_{j=1}^m \lambda_j \mathbf{h}_j \right\|_2^2, \forall \lambda \in \mathcal{C}(b, J) \right\},$$

where $b \geq 0$ is a nonnegative constant, $J \subset [m]$, and $\mathcal{C}(b, J)$ is defined as

$$\mathcal{C}(b, J) = \left\{ \lambda \in R^m : \sum_{j\notin J} \lambda_j^2 \leq b^2 \sum_{j\in J} \lambda_j^2 \right\}.$$

$\mathcal{C}(b, J)$ defines a set of sparse vector in which the components in J dominates over the other components measured by their absolute values. When $b = 0$, vectors in $\mathcal{C}(b, J)$ only have non-zero elements in set J, leading to the standard definition of sparse vectors. $\beta(b, J, H)$ essentially captures the linearly dependence among vectors in H. For instance, when all $\mathbf{h}_j$ are normalized and orthogonal to each other, we have $\beta(0, J, H) = 1$. We extend $\beta(b, J, H)$ to $\beta(d, H)$ by taking into account all the vectors with no more than d non-zero elements,

$$\beta(d, H) = \inf\{\beta(0, J, H) : J \subset [m], |J| \leq d\}.$$

We now generalize the above definitions to capture the "dependence" among the kernel matrices $\mathcal{K} = \{\widehat{K}_1, \ldots, \widehat{K}_m\}$, where $\widehat{K}_j = K_j/N$. Since we need to deal with a sparse matrix $A = (\mathbf{a}_1, \ldots, \mathbf{a}_m) \in \mathbb{R}^{N\times m}$, we extend the definition of $\mathcal{C}(b, J)$ to $\mathcal{S}(b, J, \mathcal{K})$ for sparse matrix as follows

$$\mathcal{S}(b, J, \mathcal{K}) = \tag{9}$$
$$\left\{ A = (\mathbf{a}_1, \ldots, \mathbf{a}_m) \in \mathbb{R}^{N\times m} : \sum_{j\notin J} \|\mathbf{a}_j\|_2 \leq b \sum_{j\in J} \|\mathbf{a}_j\|_2, \mathbf{a}_j \in \text{span}(K_j), j = 1, \ldots, m \right\},$$

where $\text{span}(K_j)$ stands for the subspace spanned by the column vectors of K_j. We then define quantity $\gamma(b, J, \mathcal{K})$ to capture the "dependence" among matrices in $\mathcal{K}$

$$\gamma(b, J, \mathcal{K}) = \inf \left\{ \gamma > 0 : \sum_{j\in J} \|\mathbf{a}_j\|_2 \leq \gamma \left\| \sum_{j=1}^m \widehat{K}_j \mathbf{a}_j \right\|_2, \forall A \in \mathcal{S}(b, J, \mathcal{K}) \right\}. \tag{10}$$

We finally define $\gamma(d, \mathcal{K})$ to take into account any matrix A that has no more than d non-zero column vectors

$$\gamma(d, \mathcal{K}) = \inf \left\{ \gamma(0, J, \mathcal{K}) : J \subset [m], |J| \leq d \right\}. \tag{11}$$

We note that the value of $\gamma(d, \mathcal{K})$ is closely related to the correlation between the subspace spanned by any two matrices in $\mathcal{K}$. For example, when subspaces spanned by each matrix

$\widehat{K}_j$ are orthogonal to each other and let the minimum non-zero eigenvalues of $\widehat{K}_j, j \in [m]$ be larger than $\sigma^+_{\min} \leq 1$, we have $\gamma(d, \mathcal{K}) \leq \sqrt{d}/\sigma^+_{\min}$. More generally, if we let $\delta(\mathcal{K})$ denote the correlation between the subspace spanned by any two matrices in $\mathcal{K}$, defined as

$$\delta(\mathcal{K}) = \max_{1 \leq i < j \leq d} \max_{\mathbf{a}_i, \mathbf{a}_j} \frac{|(\widehat{K}_i \mathbf{a}_i)^\top (\widehat{K}_j \mathbf{a}_j)|}{\|\widehat{K}_i \mathbf{a}_i\|_2 \|\widehat{K}_j \mathbf{a}_j\|_2}.$$

The following proposition shows the relationship between $\gamma(d, \mathcal{K})$ and $\delta(\mathcal{K})$ when $\delta(\mathcal{K})$ is small.

Proposition 2 *If $\delta(\mathcal{K}) < \dfrac{1}{d-1}$, the following inequality holds for $\gamma(d, \mathcal{K})$ and $\delta(\mathcal{K})$,*

$$\gamma(d, \mathcal{K}) \leq \frac{\sqrt{d}}{\sqrt{1 - (d-1)\delta(\mathcal{K})}\sigma^+_{\min}},$$

where $\sigma^+_{\min}$ is a lower bound of the minimum non-zero eigenvalues of $\widehat{K}_j, j \in [m]$.

Remark: The correlation between different kernels has beed used in the previous studies for proving learning bounds for multiple kernel learning. For example, in (Cortes et al., 2009), the authors derived generalization bounds for kernel ridge regression with ℓ_2 regularization on multiple kernels in the case where the kernels are orthogonal.

The following lemma shows that when $\gamma(2d, \mathcal{K})$ is bounded, the solution f of the Algorithm 2 converges to f^* in a geometric rate.

Lemma 3 *Let f be the solution output from Algorithm 2, and $(f^*, \hat{f}, \varepsilon^*)$ be defined in (8). For any $\mu \geq 1$, we have either $\mathcal{E}_N(f) - \mathcal{E}_N(f^*) \leq \mu(\mathcal{E}_N(\hat{f}) - \mathcal{E}_N(f^*))$ or*

$$\mathcal{E}_N(f) - \mathcal{E}_N(f^*) \leq \frac{1}{2}\left[\max(0, 1-\tau)\right]^d,$$

where τ is defined as

$$\tau = \frac{(\mu-1)^2}{8\mu(\mu+1)\gamma(2d, \mathcal{K})}.$$

The proof is deferred to Appendix B.

As indicated by Lemma 3, Algorithm 2 achieves a geometric convergence rate of $(1-\tau)^d$, where τ depends on the parameter $\gamma(2d, \mathcal{K})$. In particular, the smaller the $\gamma(2d, \mathcal{K})$, the faster the convergence. One shortcoming with Lemma 3 is that it does not give the explicit expression for bounding $\mathcal{E}_N(f) - \mathcal{E}_N(f^*)$ because the bound depends on parameter μ. The following theorem makes the bound more explicit.

Theorem 4 *Let f be the solution output from Algorithm 2, and (f^*, ε^*) be defined in (8). If the number of selected kernels d is sufficiently large, i.e.,*

$$d \geq 16\gamma(2d, \mathcal{K}) \ln\left(\frac{1}{12\varepsilon^*}\right),$$

then we have

$$\mathcal{E}_N(f) - \mathcal{E}_N(f^*) \leq 6\varepsilon^*.$$

Proof According to Lemma 3, we have

$$\mathcal{E}_N(f) - \mathcal{E}_N(f^*) \leq \min_{\mu \geq 1} \max \left(\mu \varepsilon_*, \frac{1}{2} \left[\max(0, 1-\tau) \right]^d \right).$$

It is straightforward to show that for any $z \in [0,1)$, if $\mu \geq (2+z)/(1-z)$, we have $\tau > z/[8\gamma]$, where $\gamma = \gamma(2d, \mathcal{K})$. We thus have

$$\mathcal{E}_N(f) - \mathcal{E}_N(f^*) \leq \min_{z \in [0,1)} \max \left(\frac{3\varepsilon_*}{1-z}, \frac{1}{2} \exp(-dz/[8\gamma]) \right) \leq \min_{z \in [0,1)} \max \left(\frac{3\varepsilon_*}{1-z}, \frac{1}{2} \exp\left(2z \ln(12\varepsilon^*)\right) \right).$$

The optimum of R.H.S is achieved when

$$\frac{3\varepsilon_*}{1-z} = \frac{1}{2} \exp\left(2z \ln(12\varepsilon^*)\right).$$

Under the condition given in the theorem, we have the above equation satisfied if $z = 1/2$. We also note that the solution to the above equation is unique because $\frac{3\varepsilon_*}{1-z} - \frac{1}{2} \exp\left(2z \ln(12\varepsilon^*)\right)$ is monotonically increasing in z. We complete the proof by plugging $z = 1/2$. ■

4.2. Generalization Bound

As previously mentioned, there is a rich body of literature dealing with the generalization error bounds of MKL algorithms (Hussain and Shawe-Taylor, 2011; Ying and Zhou, 2007; Bousquet and Herrmann, 2003; Srebro and Ben-david, 2006; Ying and Campbell, 2009). In the remarkable work of (Lanckriet et al., 2004), a convergence rate of $O(\sqrt{m/N})$ has been proved for MKL with ℓ_1 constraint. After that, this bound is improved utilizing the pseudo-dimension of the given kernel class in (Srebro and Ben-david, 2006). Cortes et al. (2009) studied the problem of multiple kernel learning with ℓ_2 regularization for regression, and derived learning bounds that have an additive term $O(\sqrt{m/N})$ when kernels are orthogonal. In (Cortes et al., 2010) new generalization bounds for the family of convex combination of kernel function with ℓ_1 constraint were presented which have logarithmic dependency on the number of kernels (i.e., $\sqrt{\ln m}$). It is worth mentioning that although the mentioned generalization bounds differ in their dependency on the number of base kernels, however, all convergence rate presented are of order $1/\sqrt{N}$ with respect to the number N of samples. It is worth mentioning that although the mentioned generalization bounds differ in their dependency on the number of base kernels, however, all convergence rate presented are of order $1/\sqrt{N}$ with respect to the number N of samples. Recently, (Kloft and Blanchard, 2011) utilized local Rademacher complexity and derived a tighter upper bound with respect to N for ℓ_p norm MKL by considering the decay rate of eigenvalues of kernel matrices. Suzuki (2011) presented a unified framework to derive the bounds of MKL with arbitrary mixed-norm type regularization.

To present the generalization error bound for the sparse MKL solution obtained by Algorithm 2, we introduce the following bounded RKHS $\mathcal{H}(R)$ as

$$\mathcal{H}(R) = \left\{ f = \sum_{j=1}^{m} f_j : f_j \in \mathcal{H}_j, j \in [m], \sum_{j=1}^{m} \|f_j\|_{\mathcal{H}_j} \leq R \right\}.$$

The generalization error bound is stated in the following theorem.

Theorem 5 *Let f be the solution output from Algorithm 2, (f^*, ε^*) be defined in (8),and f_R^* be the optimal function for minimizing the expected loss in $\mathcal{H}(R)$, i.e. $f_R^* = \arg\min_{f\in\mathcal{H}(R)} \mathcal{E}(f)$. Assuming $A > 1$, $m \geq 3$, and $A\ln(m+1) \leq N \leq 2^{m+1}$, we have either $\|f - f_R^*\| \leq 8\max(R, \sqrt{d})/\sqrt{N}$ or with a probability at least $1-(m+1)^{-A+1}$,*

$$\mathcal{E}(f) - \mathcal{E}(f_R^*) \leq \mathcal{E}_N(f) - \mathcal{E}_N(f^*) + 196(R+\sqrt{d})^2\sqrt{\frac{A\ln(m+1)}{N}}.$$

Under the assumption $d \geq 16\gamma(2d, \mathcal{K})\ln\left(\frac{1}{12\varepsilon^*}\right)$, we have

$$\mathcal{E}(f) - \mathcal{E}(f_R^*) \leq 6\varepsilon^* + 196(R+\sqrt{d})^2\sqrt{\frac{A\ln(m+1)}{N}}.$$

Remark: First, we should note that there is a tradeoff in the generalization bound with respect to d, since ε^* could increase when d decreases. Second, the generalization bound of the proposed algorithm for learning a combination of no more than d kernels has an additive term $O(d\sqrt{\ln m/N})$, which deteriorates by a factor of d compared to previous learning bounds of MKL. Third, if we assume ε^* is small, e.g., in the order of $O(N^{-1/2})$, and $\gamma(2d, \mathcal{K}) \leq O(\sqrt{d})$, we can let $d = O(\ln^2 N)$, i.e. learning a combination of no more than $O(\ln^2 N)$ kernels, and we have the generalization error of the proposed algorithm bounded by $O(\ln^2 N\sqrt{\ln m/N})$, which only deteriorates by a factor of $\ln^2 N$ compared with the best known learning bound of MKL (i.e. $O(\sqrt{\ln m/N})$).

In order to prove Theorem 5, we need the following lemma to bound the concentration of regression error, where $(\ell \circ f)(\mathbf{x}, y) = \ell(f(\mathbf{x}), y)$, and P_N and P are defined by

$$P_N(F) \quad = \quad \frac{1}{N}\sum_{i=1}^{N} F(\mathbf{x}_i, y_i), \quad P(F) = \mathrm{E}_{\mathbf{x},y}[F(\mathbf{x}, y)],$$

for any function F that takes $(\mathbf{x}, y)$ as input.

Lemma 6 *Define $r_0 = 8R/\sqrt{N}$ and $L = R+1$. Let $g \in \mathcal{H}(R)$ be a fixed function. Assume $A > 1$, and $A\ln(m+1) \leq N \leq 2^{m+1}$. With a probability at least $1-(m+1)^{-A+1}$, for any $f \in \mathcal{H}(R)$, and any $r > r_0$, we have*

$$\sup_{\sum_{i=1}^m \|f_i - g_i\|_{\mathcal{H}_j} \leq r} |(P - P_N)(\ell \circ f - \ell \circ g)| \leq 88Lr\sqrt{\frac{A\ln(m+1)}{N}}.$$

The proof of Lemma 6 is provided in Appendix D. We are now ready to prove Theorem 5.

Proof [of Theorem 5] First, we show that the solution f obtained by Algorithm 2 has a bounded functional norm $\|f\|$. We have

$$\|f\| = \left\|\sum_{k=1}^{d} f_{j_k}\right\| \leq \sum_{k=1}^{d} \|f_{j_k}\|_{\mathcal{H}_{j_k}}.$$

Following inequality (14) in the Proof of Lemma 3, we have

$$\|f_{j_k}\|_{\mathcal{H}_{j_k}} = \frac{1}{N^2}\mathbf{a}^k K_{j_k}\mathbf{a}^k = \|\nabla_{j_k}\mathcal{E}_N(f^k)\|_{\mathcal{H}_{j_k}}.$$

According to the inequality in (13) in the Proof of Lemma 3, we have

$$\|f_{j_k}\|^2_{\mathcal{H}_{j_k}} \leq 2\left(\mathcal{E}_N(f^{k-1}) - \mathcal{E}_N(f^k)\right),$$

due to $\|\nabla_{j_k}\mathcal{E}_N(f^k)\|_{\ell_2(\mathcal{D})} \leq \|\nabla_{j_k}\mathcal{E}_N(f^k)\|_{\mathcal{H}_{j_k}}$. Hence

$$\|f\| \leq \sum_{k=1}^{d}\|f_{j_k}\|_{\mathcal{H}_{j_k}} \leq \sqrt{d}\sqrt{\sum_{k=1}^{d}\|f_{j_k}\|^2_{\mathcal{H}_{j_k}}} \leq \sqrt{2d\mathcal{E}_N(f^0)} \leq \sqrt{\frac{d}{N}}\|\mathbf{y}\|_2 \leq \sqrt{d}.$$

Second, we have

$$\begin{aligned}
\mathcal{E}(f) &\leq \mathcal{E}(f_R^*) + \mathcal{E}_N(f) - \mathcal{E}_N(f_R^*) + \mathcal{E}(f) - \mathcal{E}_N(f) + \mathcal{E}_N(f_R^*) - \mathcal{E}(f_R^*) \\
&\leq \mathcal{E}(f_R^*) + \mathcal{E}_N(f) - \mathcal{E}_N(f_R^*) + \sup_{f\in\mathcal{H}(\sqrt{d})}|(P - P_N)(\ell\circ f - \ell\circ f_R^*)| \\
&\leq \mathcal{E}(f_R^*) + \mathcal{E}_N(f) - \mathcal{E}_N(f^*) + \sup_{f\in\mathcal{H}(\sqrt{d})}|(P - P_N)(f - f_R^*)|.
\end{aligned}$$

Using the Lemma 6, we have either $\|f - f_R^*\| \leq 8\max(R, \sqrt{d})/\sqrt{N}$, or with a probability at least $1 - (m+1)^{-A+1}$, that

$$\begin{aligned}
\sup_{f\in\mathcal{H}(\sqrt{d})}|(P - P_N)(\ell\circ f - \ell\circ f_R^*)| &\leq \sup_{\|f-g\|\leq R+\sqrt{d}}|(P - P_N)(\ell\circ f - \ell\circ g)| \\
&\leq 88(\max(R,\sqrt{d}) + 1)(R + \sqrt{d})\sqrt{\frac{A\ln(m+1)}{N}} \\
&\leq 196(R + \sqrt{d})^2\sqrt{\frac{A\ln(m+1)}{N}},
\end{aligned}$$

leading to

$$\mathcal{E}(f) \leq \mathcal{E}(f_R^*) + \mathcal{E}_N(f) - \mathcal{E}_N(f^*) + 196(R + \sqrt{d})^2\sqrt{\frac{A\ln(m+1)}{N}}.$$

We complete the proof by plugging the result from Theorem 4. ■

5. Conclusion

In this paper, we developed an efficient algorithm for sparse multiple kernel learning (MKL) based on greedy coordinate descent algorithm. By using an empirical ℓ_2 norm for measuring the size of functional gradients, we are able to achieve a geometric convergence rate under certain conditions. We also prove the generalization error bound of the proposed algorithm. As the future work, we plan to provide better quantization about the independence among kernel matrices, a key condition for our algorithm to achieve geometric convergence.

References

Andreas Argyriou, Charles A. Micchelli, and Massimiliano Pontil. Learning convex combinations of continuously parameterized basic kernels. In *Proceedings of the 18th Annual Conference on Learning Theory*, pages 338–352, 2005.

Andreas Argyriou, Raphael Hauser, Charles A. Micchelli, and Massimiliano Pontil. A dc-programming algorithm for kernel selection. In *Proceedings of the 23rd international conference on Machine learning*, pages 41–48, 2006.

Francis Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems*, pages 105–112, 2008.

Francis R. Bach, Gert R. G. Lanckriet, and Michael I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the 21st International Conference on Machine learning*, pages 6–13, 2004.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, pages 44–58, 2002.

Olivier Bousquet and Daniel J. L. Herrmann. On the complexity of learning the kernel matrix. In *Proceedings of the 17th Annual Conference on Neural Information Processing Systems*, pages 399–406, 2003.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. L2 regularization for learning kernels. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 109–116, 2009.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Generalization bounds for learning kernels. In *Proceedings of the 27th Internationl Conference on Machine Learning*, 2010.

Zakria Hussain and John Shawe-Taylor. A note on improved loss bounds for multiple kernel learning. *CoRR*, abs/1106.6258, 2011.

Marius Kloft and Gilles Blanchard. The local rademacher complexity of lp-norm multiple kernel learning. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems*, 2011.

Marius Kloft, Ulf Brefeld, Soeren Sonnenburg, Pavel Laskov, Klaus-Robert Müller, and Alexander Zien. Efficient and accurate lp-norm multiple kernel learning. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, pages 997–1005. 2009.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer, 2011.

Vladimir Koltchinskii and Ming Yuan. Sparse recovery in large ensembles of kernel machines on-line learning and bandits. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 229–238, 2008.

Vladimir Koltchinskii and Ming Yuan. Sparsity in multiple kernel learning. *Annuals of Statistics*, 38:3660–3694, 2010.

Gert R. G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, December 2004.

Darrin P. Lewis, Tony Jebara, and William Stafford Noble. Nonstationary kernel combination. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 553–560, 2006.

Charles A. Micchelli and Massimiliano Pontil. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6:1099–1125, 2005.

Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. CORE Discussion Paper #2010-2, 2010.

Cheng Soon Ong, Alexander J. Smola, and Robert C. Williamson. Learning the kernel with hyperkernels. *Journal of Machine Learning Research*, 6:1043–1071, December 2005.

Francesco Orabona and Luo Jie. Ultra-fast optimization algorithm for sparse multi kernel learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 249–256, 2011.

A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet. Simplemkl. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

Shai Shalev-Shwartz, Nathan Srebro, and Tong Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, 20(6): 2807–2832, 2010.

Vikas Sindhwani and Aurelie C. Lozano. Non-parametric group orthogonal matching pursuit for sparse learning with multiple kerenels. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems*, 2011.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153–172, 2007.

Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006.

Nathan Srebro and Shai Ben-david. Learning bounds for support vector machines with learned kernels. In *Proceedings of the 19th Annual Conference on Learning Theory*, pages 169–183, 2006.

Taiji Suzuki. Unifying framework for fast learning rate of non-sparse multiple kernel learning. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems*, 2011.

Taiji Suzuki and Ryota Tomioka. SpicyMKL: a fast algorithm for multiple kernel learning with thousands of kernels. *Machine Learning*, pages 1–32, 2011.

S. V. N. Vishwanathan, Zhaonan sun, Nawanol Ampornpunt, and Manik Varma. Multiple kernel learning and the smo algorithm. In *Proceedings of the 24th Annual Conference on Neural Information Processing Systems*, pages 2361–2369, 2010.

Z. Xu, R. Jin, I. King, and M. R. Lyu. An extended level method for efficient multiple kernel learning. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems*, pages 1825–1832, 2008.

Yiming Ying and Colin Campbell. Generalization bounds for learning the kernel. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Yiming Ying and Ding-Xuan Zhou. Learnability of gaussians with flexible variances. *Journal of Machine Learning Research*, 8, December 2007.

Sangwoon Yun, Paul Tseng, and Kim-Chuan Toh. A block coordinate gradient descent method for regularized convex separable optimization and covariance selection. *Mathematical Programming*, 129(2):331–355, 2011.

Appendix A. [Proof of Theorem 1]

First, we bound the difference between $\mathcal{L}(f^k)$ and $\mathcal{L}(f^*)$ and show that for $k \geq 1$, the following holds

$$\mathcal{L}(f^{k+1}) - \mathcal{L}(f^*) \leq \frac{2\|f^*\|^2}{k}. \tag{12}$$

Similar to the standard theory of greedy algorithm (Shalev-Shwartz et al., 2010), we have

$$\mathcal{L}(f^k) - \mathcal{L}(f^*) \leq \sum_{j=1}^{m} \left\langle f_j^k - f_j^*, \nabla_j \mathcal{E}_N(f^k) + \lambda \delta_j \right\rangle_{\mathcal{H}_j},$$

where $\delta_j \in \partial_j \|f_j^k\|_{\mathcal{H}_j}$. Since f^k is the optimal solution of $\mathcal{E}_N(f) + \lambda\|f\|$ on the support $J(f^k)$, we have $\nabla_j \mathcal{E}_N(f^k) + \lambda \partial_j \|f_j^k\|_{\mathcal{H}_j} = 0, \forall j \in J(f^k)$. By choosing

$$\delta_j = -\frac{\nabla_j \mathcal{E}_N(f^k)}{\max(\lambda, \|\nabla_j \mathcal{E}_N(f^k)\|_{\mathcal{H}_j})}, j \notin J(f^k),$$

we have

$$\mathcal{L}(f^k) - \mathcal{L}(f^*) \quad \leq \quad \sum_{j \notin J(f^k)} \left\langle -f_j^*, \nabla_j \mathcal{E}_N(f^k) + \lambda \delta_j \right\rangle_{\mathcal{H}_j} \leq \|f^*\| \left[\max_{j \in [m]} |\nabla_j \mathcal{E}_N(f^k)|_{\mathcal{H}_j} - \lambda \right]_+,$$

where $[z]_+ = \max(0, z)$. The above inequality indicates that if $\max_{j \in [m]} \|\nabla_j \mathcal{E}_N(f^k)\|_{\mathcal{H}_j} \leq \lambda$, f^k is the optimal solution, we thus exist the loop.

In the following, we assume $\max_{j\in[m]} \|\nabla_j \mathcal{E}_N(f^k)\|_{\mathcal{H}_j} > \lambda$. We have

$$\begin{aligned}\mathcal{L}(f^{k+1}) &= \min_{J(f)=\mathcal{S}_{k+1}} \mathcal{E}_N(f) + \lambda\|f\| \\ &\leq \min_{J(f)=\mathcal{S}_{k+1}} \mathcal{E}_N(f^k) + \lambda\|f\| + \sum_{j=1}^{m} \left\langle f_j - f_j^k, \nabla_j \mathcal{E}_N(f^k) \right\rangle_{\mathcal{H}_j} + \frac{1}{2N}\sum_{i=1}^{N}(f(\mathbf{x}_i) - f^k(\mathbf{x}_i))^2,\end{aligned}$$

where the inequality follows the definition of $\mathcal{E}_N(f)$. To bound the R.H.S., we consider the following construction of f

$$f = f^k - \eta g_{j_{k+1}} = f^k - \eta \frac{\nabla_{j_{k+1}} \mathcal{E}_N(f^k)}{\|\nabla_{j_{k+1}} \mathcal{E}_N(f^k)\|_{\mathcal{H}_j}}.$$

Using the above solution f, we have

$$\mathcal{L}(f^{k+1}) \leq \mathcal{L}(f^k) + \eta\lambda - \eta\|\nabla_{j_{k+1}} \mathcal{E}_N(f^k)\|_{\mathcal{H}_j} + \frac{\eta^2}{2N}\sum_{i=1}^{N}[g_{j_{k+1}}(\mathbf{x}_i)]^2.$$

Since the above inequality hold for any $\eta \geq 0$ and $j_{k+1} = \arg\max_j \|\nabla_j \mathcal{E}_N(f^k)\|$, we have

$$\begin{aligned}\mathcal{L}(f^{k+1}) &\leq \mathcal{L}(f^k) + \min_{\eta\geq 0} -\eta\left(\max_{j\in[m]} \|\nabla_j \mathcal{E}_N(f^k)\|_{\mathcal{H}_j} - \lambda\right) + \frac{\eta^2}{2N}\sum_{i=1}^{N}[g_{j_{k+1}}(\mathbf{x}_i)]^2 \\ &\leq \mathcal{L}(f^k) + \min_{\eta\geq 0} -\eta\left(\max_{j\in[m]} \|\nabla_j \mathcal{E}_N(f^k)\|_{\mathcal{H}_j} - \lambda\right) + \frac{\eta^2}{2} \\ &\leq \mathcal{L}(f^k) - \frac{1}{2}\left[\max_{j\in[m]} \|\nabla_j \mathcal{E}_N(f^k)\|_{\mathcal{H}_j} - \lambda\right]_+^2,\end{aligned}$$

where the second step follows $\|g_{j_{k+1}}\|_{\mathcal{H}_j} \leq 1$ and therefore $|g_{j_{k+1}}(\mathbf{x}_i)| \leq 1$ since $\kappa_j(\mathbf{x}_i, \mathbf{x}_i) \leq 1$. As a result, when $\max_{j\in[m]} \|\nabla_j \mathcal{E}_N(f^k)\|_{\mathcal{H}_j} - \lambda > 0$, we have

$$\mathcal{L}(f^k) - \mathcal{L}(f^{k+1}) \geq \frac{\left(\mathcal{L}(f^k) - \mathcal{L}(f^*)\right)^2}{2\|f^*\|^2}.$$

Define $\epsilon_k = \mathcal{L}(f^k) - \mathcal{L}(f^*)$. We have

$$\frac{1}{\epsilon_{k+1}} - \frac{1}{\epsilon_k} \geq \frac{\mathcal{L}(f^k) - \mathcal{L}(f^{k+1})}{\epsilon_k^2} \geq \frac{1}{2\|f^*\|^2},$$

leading to the result in (12).

Next, we consider two cases. In the first case, if f is obtained in the middle of the loop, we have $\max_{j\in[m]} \|\nabla_j \mathcal{E}_N(f)\|_{\mathcal{H}_j} \leq \lambda$, and therefore have $\mathcal{L}(f) = \mathcal{L}(f^*)$. If f is obtained by finishing all the loops, using (12) , we have the desired rate as

$$\mathcal{L}(f) - \mathcal{L}(f^*) \leq \frac{2\|f^*\|^2}{d-1}.$$

Appendix B. [Proof of Lemma 3]

Similar to the proof of Theorem 1, we have

$$\mathcal{E}_N(f^k) - \mathcal{E}_N(\hat{f}) \leq \sum_{j=1}^{m} \left\langle f_j^k - \hat{f}_j, \nabla_j \mathcal{E}_N(f^k) \right\rangle_{\mathcal{H}_j}.$$

According to the representer theorem, we have

$$f_j^k(\mathbf{x}) = \sum_{i=1}^{m} \alpha_{j,i}^k \kappa_j(\mathbf{x}_i, \mathbf{x}), \quad \hat{f}_j(\mathbf{x}) = \sum_{j=1}^{m} \hat{\alpha}_{j,i} \kappa_j(\mathbf{x}_i, \mathbf{x}),$$

where $\alpha_j^k = (\alpha_{j,1}^k, \ldots, \alpha_{j,n}^k)^\top \in \mathbb{R}^n$ and $\hat{\alpha}_j = (\hat{\alpha}_{j,1}, \ldots, \hat{\alpha}_{j,n})^\top \in \mathbb{R}^n$ are vector representation of function f_j^k and $\hat{f}_j$. Due to the projection step in updating the kernel classifier (step 5 in Algorithm 2), we have $\alpha_j^k \in \text{span}(K_j)$. It is also safe to assume $\hat{\alpha}_j \in \text{span}(K_j)$ because otherwise we can always project $\hat{\alpha}_j$ into the subspace $\text{span}(K_j)$ without changing the value $\hat{f}_j(\mathbf{x}_i), i \in [N]$, and therefore without change $\mathcal{E}_N(\hat{f})$. We define a norm $\|\cdot\|_a$ as

$$\|f_j^k\|_a = \sqrt{N}\|\alpha_j^k\|_2, \ \|\hat{f}_j\|_a = \sqrt{N}\|\hat{\alpha}_j\|_2.$$

Using these notations, we rewrite $\mathcal{E}_N(f^k) - \mathcal{E}_N(\hat{f})$ as

$$\begin{aligned}
\mathcal{E}_N(f^k) - \mathcal{E}_N(\hat{f}) &\leq \sum_{j=1}^{m} \left\langle f_j^k - \hat{f}_j, \nabla_j \mathcal{E}_N(f^k) \right\rangle_{\mathcal{H}_j} \leq \sum_{j=1}^{m} \|f_j^k - \hat{f}_j\|_a \left\| \nabla_j \mathcal{E}_N(f^k) \right\|_{\ell_2(\mathcal{D})} \\
&\leq \left(\sum_{j=1}^{m} \|f_j^k - \hat{f}_j\|_a \right) \max_{1 \leq j \leq m} \left\| \nabla_j \mathcal{E}_N(f^k) \right\|_{\ell_2(\mathcal{D})},
\end{aligned}$$

where the second inequality follows from Cauchy inequality and the definition of $\ell_2(\mathcal{D})$ norm of $\left\|\nabla_j \mathcal{E}_N(f^k)\right\|_{\ell_2(\mathcal{D})}$ that is given by

$$\left\| \nabla_j \mathcal{E}_N(f^k) \right\|_{\ell_2(\mathcal{D})}^2 = \frac{1}{N} \sum_{a=1}^{N} \left(\frac{1}{N} \sum_{b=1}^{N} \ell'(f^k(\mathbf{x}_b), y_b) \kappa_j(\mathbf{x}_a, \mathbf{x}_b) \right)^2 = \frac{1}{N} \|K_j \boldsymbol{\ell}'(f^k)/N\|_2^2,$$

where $\boldsymbol{\ell}'(f^k) = (\ell'(f^k(\mathbf{x}_1), y_1), \cdots, \ell'(f^k(\mathbf{x}_N), y_N))^\top$. Using the following equality

$$f^{k+1} = f^k - \frac{1}{N} \sum_{i=1}^{N} a_i^{k+1} \kappa_{j_{k+1}}(\mathbf{x}_i, \cdot),$$

where $\mathbf{a}^{k+1} = (a_1^{k+1}, \ldots, a_N^{k+1})^\top$ is the projection of vector $\boldsymbol{\ell}'(f^k)$ into the subspace $\text{span}(K_{j_{k+1}})$, we have

$$\begin{aligned}
\mathcal{E}_N(f^{k+1}) &\leq \mathcal{E}_N(f^k) + \sum_{j=1}^{m} \left\langle f_j^{k+1} - f_j^k, \nabla_j \mathcal{E}_N(f^k) \right\rangle_{\mathcal{H}_j} + \frac{1}{2N} \sum_{i=1}^{N} (f(\mathbf{x}_i) - f^k(\mathbf{x}_i))^2 \\
&= \mathcal{E}_N(f^k) - \|\nabla_{j_{k+1}} \mathcal{E}_N(f^k)\|_{\mathcal{H}_{j_{k+1}}}^2 + \frac{1}{2} \|\nabla_{j_{k+1}} \mathcal{E}_N(f^k)\|_{\ell_2(\mathcal{D})}^2 \qquad (13) \\
&\leq \mathcal{E}_N(f^k) - \frac{1}{2} \|\nabla_{j_{k+1}} \mathcal{E}_N(f^k)\|_{\ell_2(\mathcal{D})}^2,
\end{aligned}$$

where we use $f_j^{k+1} = f_j^k, \forall j \neq j_{k+1}$,

$$\begin{aligned}\left\langle f_j^{k+1} - f_j^k, \nabla_j \mathcal{E}_N(f^k) \right\rangle_{\mathcal{H}_j} &= -\frac{1}{N^2} \mathbf{a}^{k+1\top} K_{j_{k+1}} \boldsymbol{\ell}'(f^k) = -\frac{1}{N^2} \boldsymbol{\ell}'(f^k)^\top K_{j_{k+1}} \boldsymbol{\ell}'(f^k) \\ &= -\|\nabla_{j_{k+1}} \mathcal{E}_N(f^k)\|_{\mathcal{H}_j}^2. \end{aligned} \tag{14}$$

$$\frac{1}{N}\sum_{i=1}^N (f(\mathbf{x}_i) - f^k(\mathbf{x}_i))^2 = \frac{1}{N}\|K_{j_{k+1}}\mathbf{a}^k/N\|_2^2 = \frac{1}{N}\|K_{j_{k+1}}\boldsymbol{\ell}'(f^k)/N\|_2^2 = \left\|\nabla_{j_{k+1}}\mathcal{E}_N(f^k)\right\|_{\ell_2(\mathcal{D})}^2,$$

and the fact $\|\nabla_j \mathcal{E}_N(f^k)\|_{\ell_2(\mathcal{D})} \leq \|\nabla_j \mathcal{E}_N(f^k)\|_{\mathcal{H}_j}$. As a result, we have

$$\mathcal{E}_N(f^k) - \mathcal{E}_N(f^{k+1}) \geq \frac{\left(\mathcal{E}_N(f^k) - \mathcal{E}_N(\hat{f})\right)^2}{2\left(\sum_{j=1}^m \|\hat{f}_j - f_j^k\|_a\right)^2}.$$

Define $\delta_j = \alpha_j^k - \hat{\alpha}_j, j \in [m]$. Since $\alpha_j^k \in \text{span}(K_j)$ and $\hat{\alpha}_j \in \text{span}(K_j)$, we have $\delta_j \in \text{span}(K_j)$. Since we assume $\hat{f}$ is a combination of no more than d kernel classifiers, there are at most $2d$ non-zero vectors in the set $\{\delta_1, \ldots, \delta_m\}$. Using the definition of $\gamma(d, \mathcal{K})$, we have

$$\sum_{j=1}^m \|f_j^k - \hat{f}_j\|_a = \sqrt{N}\sum_{j=1}^m \|\delta_j\|_2 \leq \frac{\gamma(2d,\mathcal{K})}{\sqrt{N}} \left\|\sum_{j=1}^m K_j(\alpha_j^k - \hat{\alpha}_j)\right\|_2 = \gamma(2d,\mathcal{K})\|f^k - \hat{f}\|_{\ell_2(\mathcal{D})}.$$

To simplify our notation, we define $\gamma = \gamma(2d, \mathcal{K})$. We have

$$\begin{aligned}\mathcal{E}_N(f^k) - \mathcal{E}_N(f^{k+1}) &\geq \frac{(\mathcal{E}_N(f^k) - \mathcal{E}_N(\hat{f}))^2}{2\gamma\|f^k - \hat{f}\|_{\ell_2(\mathcal{D})}^2} \geq \frac{(\mathcal{E}_N(f^k) - \mathcal{E}_N(\hat{f}))^2}{4\gamma\left(\|f^k - f^*\|_{L_2}^2 + \|\hat{f} - f^*\|_{\ell_2(\mathcal{D})}^2\right)} \\ &\geq \frac{(\mathcal{E}_N(f^k) - \mathcal{E}_N(\hat{f}))^2}{8\gamma\left(\mathcal{E}_N(f^k) - \mathcal{E}_N(f^*) + \mathcal{E}_N(\hat{f}) - \mathcal{E}_N(f^*)\right)}.\end{aligned}$$

The last step in the above inequality follows the fact that f^* is the minimizer of the empirical loss $\mathcal{E}_N(f)$ and therefore

$$\mathcal{E}_N(f) - \mathcal{E}_N(f^*) \geq \frac{1}{2N}\sum_{i=1}^N (f(\mathbf{x}_i) - f^*(\mathbf{x}_i))^2 = \frac{1}{2}\|f - f^*\|_{\ell_2(\mathcal{D})}^2.$$

Let $k(\mu)$ be the iteration index such that for any $k \leq k(\mu)$ we have $\mathcal{E}_N(f^k) - \mathcal{E}_N(f^*) \geq \mu(\mathcal{E}_N(\hat{f}) - \mathcal{E}_N(f^*)) = \mu\varepsilon_*$, where $\mu \geq 1$. Then, for all $k \leq k(\mu)$, we have

$$\mathcal{E}_N(f^k) - \mathcal{E}_N(f^{k+1}) \geq \frac{(\mu-1)^2}{8\gamma\mu(\mu+1)}\left(\mathcal{E}_N(f^k) - \mathcal{E}_N(f^*)\right).$$

Define $\epsilon_k = \mathcal{E}_N(f^k) - \mathcal{E}_N(\hat{f})$ and $\tau = \frac{(\mu-1)^2}{8\gamma\mu(\mu+1)}$. Then, for any $k \leq k(\mu)$, we have $\epsilon_{k+1} \leq \max(0, 1-\tau)\epsilon_k$ and therefore

$$\epsilon_k \leq [\max(0,1-\tau)]^k \epsilon_0 = [\max(0,1-\tau)]^k \frac{\|\mathbf{y}\|_2^2}{2N} \leq \frac{1}{2}[\max(0,1-\tau)]^k,$$

leading to the result in the lemma.

Appendix C. [Proof of Proposition 2]

We only need to prove $\widehat{\gamma} = \frac{\sqrt{d}}{\sqrt{1-(d-1)\delta(\mathcal{K})}\sigma^+_{\min}}$ satisfies the following inequality

$$\sum_{j\in J} \|\mathbf{a}_j\|_2 \leq \widehat{\gamma} \left\| \sum_{j\in J} \widehat{K}_j \mathbf{a}_j \right\|_2 .$$

To prove this, we let $\mathbf{z} = (\|\mathbf{a}_j\|_2, j \in J)$, and proceed as follows:

$$\begin{aligned}
\widehat{\gamma}^2 \left\| \sum_{j\in J} \widehat{K}_j \mathbf{a}_j \right\|_2^2 &\geq \widehat{\gamma}^2 \left(\sum_{j\in J} \|\widehat{K}_j \mathbf{a}_j\|_2^2 + \sum_{i\neq j, i,j\in J} \langle \widehat{K}_i \mathbf{a}_i, \widehat{K}_j \mathbf{a}_j \rangle \right) \\
&\geq \widehat{\gamma}^2 (\sigma^+_{\min})^2 \left(\sum_{j\in J} \|\mathbf{a}_j\|_2^2 - \delta(\mathcal{K}) \sum_{i\neq j, i,j\in J} \|\mathbf{a}_i\|_2 \|\mathbf{a}_j\|_2 \right) \\
&\geq \widehat{\gamma}^2 (\sigma^+_{\min})^2 \left((1-\delta(\mathcal{K})) \sum_{j\in J} \|\mathbf{a}_j\|_2^2 + \delta(\mathcal{K})(2\|\mathbf{z}\|_2^2 - (\mathbf{z}^\top \mathbf{1})^2) \right) \\
&\geq \widehat{\gamma}^2 (\sigma^+_{\min})^2 \left(\sum_{j\in J} \|\mathbf{a}_j\|_2^2 \right) (1-(d-1)\delta(\mathcal{K})) \geq \widehat{\gamma}^2 \frac{(\sigma^+_{\min})^2}{d} \left(\sum_{j\in J} \|\mathbf{a}_j\|_2 \right)^2 (1-(d-1)\delta(\mathcal{K})).
\end{aligned}$$

Plugging the values of $\widehat{\gamma}$, we prove the required inequality.

Appendix D. [Proof of Lemma 6]

We first bound the concentration of regression error for fixed r. Using the Telagrand inequality (Koltchinskii, 2011), we have with a probability $1 - e^{-t}$

$$\begin{aligned}
&\sup_{\|f-g\|\leq r} |(P - P_N)(\ell \circ f - \ell \circ g)| \\
&\leq 2 \left(\mathrm{E} \left[\sup_{\|f-g\|\leq r} |(P - P_N)(\ell \circ f - \ell \circ g)| \right] + \sqrt{P(\ell \circ f - \ell \circ g)^2} \sqrt{\frac{t}{N}} + |\ell \circ f - \ell \circ g|_\infty \frac{t}{N} \right) \\
&\leq 2 \left(\mathrm{E} \left[\sup_{\|f-g\|\leq r} |(P - P_N)(\ell \circ f - \ell \circ g)| \right] + Lr \sqrt{\frac{t}{N}} + \frac{Lrt}{N} \right).
\end{aligned}$$

We now bound the expectation $\mathrm{E}\left[\sup_{\|f-g\|\leq r} |(P - P_N)(\ell \circ f - \ell \circ g)| \right]$. We have

$$\begin{aligned}
&\mathrm{E} \left[\sup_{\|f-g\|\leq r} |(P - P_N)(\ell \circ f - \ell \circ g)| \right] \\
&\leq 2\mathrm{E}_{N,\sigma} \left[\sup_{\|f-g\|\leq r} R_n(\ell \circ f - \ell \circ g) \right] \leq 4L\mathrm{E}_{N,\sigma} \left[\sup_{\|f-g\|\leq r} R_N(f-g) \right],
\end{aligned}$$

where $R_N(f) = \frac{1}{N}\sum_{i=1}^N \sigma_i f(\mathbf{x}_i)$ is the Rademacher complexity measure and $\sigma_i, i = 1, \ldots, N$ are Rademacher variables. The last inequality follows the contraction property of Rademacher complexity measure (Koltchinskii, 2011). To continue bounding the quantity, we first notice that

$$\sup_{\|f-g\|\le r} R_N(f-g) \le r \max_{1\le j\le m}\left[\sup_{\|f_j-g_j\|_{\mathcal{H}_j}\le 1} R_N(f_j - g_j)\right].$$

This is because

$$\begin{aligned}
\sup_{\|f-g\|\le r} R_N(f-g) &= r \sup_{\sum_{j=1}^m \|f_j-g_j\|_{\mathcal{H}_j}\le 1} \frac{1}{N}\sum_{i=1}^N\sum_{j=1}^m \sigma_i(f_j(\mathbf{x}_i') - g_j(\mathbf{x}_i')) \\
&= r \sup_{\sum_{j=1}^m \|f_j-g_j\|_{\mathcal{H}_j}\le 1} \sum_{j=1}^m \frac{\|f_j-g_j\|_{\mathcal{H}_j}}{N}\sum_{i=1}^N \sigma_i \frac{f_j(\mathbf{x}_i') - g_j(\mathbf{x}_i')}{\|f_j-g_j\|_{\mathcal{H}_j}} \\
&\le r \max_{1\le j\le m} \sup_{\|f_j-g_j\|_{\mathcal{H}_j}\le 1} R_N(f_j-g_j).
\end{aligned}$$

Using Theorem 5 from (Hussain and Shawe-Taylor, 2011), we have, with a probability $1 - e^{-t}$, that

$$\begin{aligned}
\mathrm{E}_\sigma\left[\max_{1\le j\le m}\sup_{\|f_j-g_j\|_{\mathcal{H}_j}\le 1} R_N(f_j-g_j)\right] &\le \max_{1\le j\le m}\mathrm{E}_\sigma\left[\sup_{\|f_j-g_j\|_{\mathcal{H}_j}\le 1} R_N(f_j-g_j)\right] + 4\sqrt{\frac{\ln(m+1)+t}{2N}} \\
&\le \frac{1}{\sqrt{N}} + 4\sqrt{\frac{\ln(m+1)+t}{2N}},
\end{aligned}$$

where the last step uses the fact $\kappa_j(\mathbf{x},\mathbf{x}) \le 1$ and the result from (Bartlett et al., 2002). Combining the above results and setting $t = A\ln(m+1)$, we have with a probability at least $1 - 2(m+1)^A$, for a fixed r,

$$\begin{aligned}
&\sup_{\|f-g\|\le r} |(P - P_N)(\ell\circ f - \ell\circ g)| \\
&\le 2Lr\left(\frac{4}{\sqrt{N}} + 16\sqrt{\frac{(A+1)\ln(m+1)}{2N}} + \sqrt{\frac{A\ln(m+1)}{N}} + \frac{A\ln(m+1)}{N}\right) \\
&\le Lr\left(42\sqrt{\frac{A\ln(m+1)}{N}} + 2\frac{A\ln(m+1)}{N}\right). \qquad (15)
\end{aligned}$$

Now, we show the bound holds uniformly for all $r \in (r_0, 2R)$. Note that r cannot be larger than $2R$ because $\sum_{i=1}^m \|f_i - g_i\|_{\mathcal{H}_i} \le 2R$. To this end, we consider $R_j = 2^{1-j}R, j = 0, \ldots, j_0$, where $j_0 \le \lceil \log_2[2R] - \log_2 r_0\rceil \le 0.5\log_2 N - 1$. Then, with probability $1 - [\log_2 N](m+1)^{-A}$, we have (15) hold for all $\{R_j\}_{j=0}^{j_0}$. Using the monotonicity with respect to r, for any $r \ge r_0$, we have

$$\sup_{\|f-g\|\le r} |(P - P_N)(\ell\circ f - \ell\circ g)| \le Lr\left(84\sqrt{\frac{A\ln(m+1)}{N}} + 4\frac{A\ln(m+1)}{N}\right) \le 88Lr\sqrt{\frac{A\ln(m+1)}{N}}.$$

We complete the proof by using the relation $\log_2 N < m+1$ and $N \ge A\ln(m+1)$.